Applied Statistics

1. Explorating a Multivariate Dataset

- Prediction problem: the model
- Curse of dimensionality
- Bias-Variance trade-off

Geometry of Data

- By columns: mean, variance, covariance, orrelation, Chebyshev, geometrical interpretation
- By rows: mean and covariance for rand. vec. $\underline{X}$, linear combination of the components of $\underline{X}$, k-linear combinations of the components of $\underline{X}$

Estimators

- Estimator for $\underline{\mu}$: properties of $\bar{\underline{X}}$
- Estimator for Σ: properties of S

Variability in a multivariate sense

- Generalized variance and total variance
- Properties of $Det(S)$ and $Tr(S)$
- Spectral decomposition
- Induced distance: why Mahalanobis'

2. Principal Component Analysis (PCA)

- Problem: find $\underline{a}$ s.t. $Var(\underline{a}^T\underline{X})$ is maximum
- Geometrical Lemma
- Principal components
- Properties of principal components
- Meaning of the PCi's loadings: $Corr(Y_i, X_k)$
- PCA on standardized variables (PCA on ρ)
- PCA on the data (PCA on S)
- Geometrical view of PCA: optimal orthonormal basis and approximation error

3. Multivariate Gaussian Distribution

- General properties of $\underline{X} \sim \mathcal{N}(\underline{\mu}, \Sigma)$
- Characterization theorem and consequences
- Gaussianity and $\underline{X}_1 \perp\!\!\!\perp \underline{X}_2$
- Gaussianity and $\underline{X}_1 | \underline{X}_2 = \underline{x}_2$
- Properties of $(\underline{X} - \underline{\mu})^T \Sigma^{-1} (\underline{X} - \underline{\mu})$
- Estimators of $\underline{\mu}$ and Σ for $\underline{X} \sim \mathcal{N}_p(\underline{\mu}, \Sigma)$
- Distribution and properties of $\bar{\underline{X}}$, S, $\hat{\Sigma}$ (and Whishart's properties)
- LLN, CLT

4. Inference about the mean vector

- large n: pivotal statistics, $CR_{1-\alpha}$, testing, p-value
- small n: Hotelling's theorem, pivotal statistics, $CR_{1-\alpha}$, testing
- $CR_{1-\alpha}$ and correlation between variables

Linear combination of the mean

- One-at-the-time $CI(\underline{\mu})$ (and testing)
- Simultaneous $CI(\underline{\mu})$: ∞ linear comb. (and testing)
- Bonferroni's method for $CI(\underline{\mu})$: finite linear comb. (and testing)
- False discovery rate (FDR)

Comparing means of gaussian distributions

- Paired data
- Repeated univariate measures

5. Multivariate Analysis of Variance

- Case $p \geq 1$, $g = 2$
  goal : inference on $\underline{\mu}_1 - \underline{\mu}_2$
- Case $p = 1$, $g \geq 1$ (ANOVA)
  goal : $H_0 : \mu_1 = \mu_2 = .. = \mu_g$ vs. $H_1 : \exists \mu_i \neq \mu_j$
  (eq.) $H_0 : \tau_1 = \tau_2 = .. = \tau_g = 0$ vs. $H_1 : \exists \tau_j \neq 0$
- Case $p \geq 1$, $g \geq 2$ (MANOVA)
  goal : $H_0 : \underline{\mu}_1 = \underline{\mu}_2 = .. = \underline{\mu}_g$ vs. $H_1 : \exists \underline{\mu}_i \neq \underline{\mu}_j$
  (eq.) $H_0 : \underline{\tau}_1 = \underline{\tau}_2 = .. = \underline{\tau}_g = 0$ vs. $H_1 : \exists \underline{\tau}_j \neq \underline{0}$
- Two-ways (M)ANOVA

6. Classification

Supervised classification

- Supervised model for classification
- Optimality criterion for δ: $ECM(\delta)$
- Optimization problem: $g = 2$, $g \geq 2$
- Optimal classifier, Bayes classifier, MLE classifier
- Special cases of Bayes classifiers: QDA, LDA
- Fisher's argument for LDA
- Evaluating a classifier by the error rate: $AER(\delta)$
- K-fold cross validation
- Support vector machines

Unsupervised classification

- Dissimilarity function (quantitative, categorical)
- Dissimilarity matrix
- Distance (/dissimilarities) between clusters (/sets)
- Hierarchical agglomerative clustering algorithm: dendrogram, cophenetic dist., CPCC, Ward's method
- Non-hierarchical methods: K-means
- Graphical: multidimensional scaling (MDS)

7. Regression

- Data driven approach: CART
- Parametric approach: linear models
- Fitting the linear model: OLS
- Coefficient of determination: R^2, R^2_{adj}
- Properties of $\hat{\underline{\beta}}$, $\hat{\underline{\epsilon}}$
- Model with $\underline{\epsilon} \sim \mathcal{N}_n(\underline{0}, \sigma^2 I)$:
  → properties of $\hat{\underline{\beta}}$, $\hat{\underline{\epsilon}}$
  → $CR_{1-\alpha}(\underline{\beta})$, $CI_{1-\alpha}(\sigma^2)$, $Sim\ CI_{1-\alpha}(\underline{a}^T\underline{\beta})$
  → $Sim\ CI_{1-\alpha}(\beta_i)$
  → Testing β's
- Prediction (Y_0, not $\mathbb{E}[Y_0|\underline{Z}_0]$)
- Generalized Least Squares (GLS)
  (special case) weighted least squares
- Diagnostic for linear models:
  → Residual analysis
  → Check for gaussianity
  → Test for autocorrelation
  → Influencial cases: leverages
  → Collinearity: VIF coefficient
- Diagnostic: collinearity and variables selection
  → Checking all possible models
  → Iterative procedures: forward/backward
  → PCA regression
  → Ridge regression
  → Lasso regression

8. Permutation Tests

- Univariate:
  → Test for 2 independent populations
  likelihood transformations: ***units permutations***
- Multivariate *(*)* [1]:
  → Test for 2 independent multivariate populations
  likelihood transf.s: ***units permutations***
  → Test for 1 multivariate population:
  center of symmetry (symmetry assumption)
  likelihood transf.s: ***units reflection on center***
  (extension to two paired multivariate populations)
- (M)ANOVA
  → One-way (M)ANOVA
  likelihood transf.s: ***lables permutations***
  (F-statistics for univariate, Wilks statistics for multivariate (*) [2]*)*
  → Two-way ANOVA
  likelihood transf.s: ***residuals permutations***
- Regression
  → Test for all the regressors (F-test)
  likelihood trans.fs: ***responses permutations*** *or* ***residuals permutations***
  → Test for one regressor (t-test)
  likelihood transf.s: ***residuals permutations***

9. Spatial Data (Geostatistics)

Spatial dependence

- Mean and covariance assumptions
- Second order stationary
- Covariogram, algebraic properties
- Variogram, algebraic and structural properties

Estimate spatial dependence

- Empirical estimate
- Model estimate: parametric family

(Spatial) Prediction

- Predictor for Z_{S_0}
- Ordinary Kriging
- Universal Kriging

10. Functional Data Analysis

- Hilbert space model for functional data
- Smoothing and interpolation of functional data:
  → basis functions
  → least square smoothing
  → smoothing with penalization
- FDA and dimensionality reduction in Hilbert spaces
- Data alignment and clustering:
  → phases and amplitude variability
  → decoupling phase and amplitude variability
  → K-mean alignment

(*) [1] In the multivariate settings, permutations and reflections are made on units, so by rows, not by columns

(*) [2] The Wilks statistic leads to rejection for small values, contrary of the F-statistic